Introduction to Survival Analysis

INTD8065 Data Analysis for Cancer Research
María-Eglée Pérez and Luis Raúl Pericchi

Contents of this class

1. Time to event data
   - Survival function and hazard function
   - Censoring

2. Kaplan-Meier estimator for the survival function
   - Confidence intervals for the survival function

Time to event data

In many clinical studies, the variable of interest is the time until something happens:

- Time of survival of a patient participating in a new drug study.
- Time from vaccination till a child suffers a rotavirus positive diarrhea.
- Lifetime of an electronic component.

All these are examples of **time to event data** or **survival data**.

Survival function and hazard function

For a time to event random variable T, the **survival function** is defined as

$$S(t) = P(T > t) = 1 - F(t)$$

Where F is the distribution function of T $(F(t) = P(T \leq t))$. Then, the survival function is the probability that the time to the event is greater than t.

Another important function is the **hazard function** or force of mortality $h(t)$, which measures the (infinitesimal) risk of dying within a short interval of time t, given that the subject is alive at time t.

$$h(t) = \frac{f(t)}{S(t)} = \frac{f(t)}{1 - F(t)}$$

Here, $f(t)$ is the density function of the variable T.

Censoring

Sometimes, it is not possible to get information about all the participants in a survival study

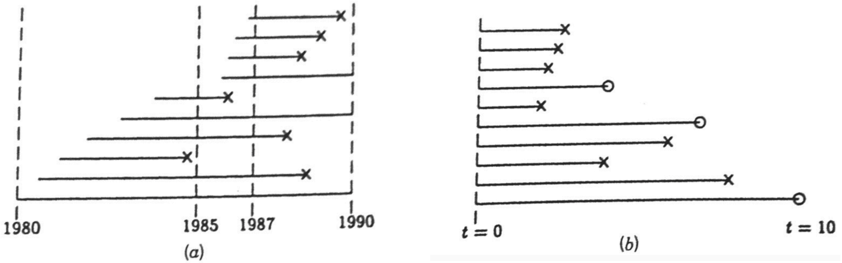
- *Loss to follow up:* the patient stop attending follow up meetings (moved to other city, etc.)
- *Dropout:* The effects of a therapy are so bad that the treatment has to be stopped.
- *Termination of the study:* (for studies with pre-determined stop dates).
- *Death due to other causes.*

When some of this situations happens, we say that our data is *censored*.

There are several types of censoring, but we will concentrate on *right censoring*.

The main assumption made to deal with this situation is that, conditional on the values of any explanatory variable, *the prognosis for any individual who has survived to a certain time t should no be affected if the individual is censored at t*. In other words, survival condition and reason of loss are independent.

Also, it can be seen that in this context the time origin need not to be (and usually is not) the same calendar time for each individual.



(a) Real timeline and (b) Time from origin (Taken from Le 1997)

Kaplan-Meier estimator for the survival function

Let $t_1 < t_2 < \ldots < t_k$ the times when events happen, and let S_{i-1} the subjects surviving until time time t_{i-1} and that have not been censored. In the period from t_{i-1} and t_i, S_i subjects survive, d_i subjects fail, and l_i subjects are censored at time t_i. Thus $S_{i-1} = S_i + d_i + l_i$.
We can estimate the probability of surviving to time t_i given that a patient has survived up to time t_{i-1} by

$$\hat{p}_i = \left(1 - \frac{d_i}{S_{i-1}}\right) = \left(1 - \frac{d_i}{S_i + d_i + l_i}\right)$$

The l_i subjects censored at time t_i do not contribute to the estimation of the survival function at time $> t_i$. However, these subjects do contribute to the estimation of the survival function at time $\leq t_i$.

Then the survival function at time t_i is estimated by

$$\begin{aligned}\hat{S}(t_i) &= \hat{p}_1 \times \hat{p}_2 \times \ldots \times \hat{p}_i \\ &= \prod_{j=1}^{i} \left(1 - \frac{d_j}{S_{j-1}}\right)\end{aligned}$$

Confidence intervals for the survival function

Construction of confidence intervals for the survival function uses the following result

$$\hat{s}^2(t) = \widehat{Var}\left[\log \hat{S}(t)\right] = \sum_{j=1}^{i} \frac{d_j}{S_{j-1}(S_{j-1} - d_j)}$$

The $(1-\alpha) \times 100\%$ confidence interval is then

$$(\hat{S}(t) \times \exp(-z_{\alpha/2}\hat{s}(t)), \hat{S}(t) \times \exp(z_{\alpha/2}\hat{s}(t)))$$

Calculations in R use the package `survival` developed by Terry Therneau and ported to R by Thomas Lumley. This package is included in the standard distribution of R.

Package `survival` also allows to compare survival curves using the *log-rank test*.